A Simple Method for Estimating the Joint Failure Time and Failure Mileage Distribution from Automobile Warranty Data

Tim P. Davis

About the Author

November 12, 1999 http://www.rlis.ford.com/ftj/ Vol. 2, Issue 6

Abstract

Lawless, Hu, and Cao (1995) present a method for the analysis of the important problem of estimation of survival rates from automobile warranty data when both time to failure and mileage to failure are of interest. In their paper, they choose to model, marginally, the distribution of mileage to failure, and then, conditionally, the distribution of time to failure, given mileage. In this short article, we present an alternative approach to the problem, which can in some cases be simpler, and illustrate it with the analysis of a real problem.

1. Introduction

Lawless, Hu, and Cao (1995 - henceforth LHC) present a method for estimating the joint density of time to failure, T, and mileage to failure, M, from some automobile warranty data. In order to keep this article as concise as possible, the reader is referred to LHC for definitions, terminology, and references for key results in reliability and failure time analysis.

Essentially the approach of LHC is to model $f_M(m)$, the marginal density of M, and also $f_{T|M}(t|m)$, the conditional density of T given M, and then derive the joint density of T and M as $f_{T,M}(t,m) = f_{T|M}(t|m)\, f_M(m)$. This is confirmed by equation (2.4) of LHC, although there the focus is on the mileage accumulation rate across the population of drivers, which LHC denote with their variable U. The idea is to avoid estimating $f_{T,M}(t,m)$ directly, because of potentially complicated censoring mechanisms involving time and mileage restrictions typical in automobile warranty, and almost certain dependence between T and M. Censoring is discussed in some detail in Section 3.

However, as LHC concede, even estimating $f_{T|M}(t|m)$ and $f_M(m)$ is not straightforward - information is needed on the accumulation of mileage not only for the failed specimens (which are usually known from the appropriate warranty claim) but also on unfailed specimens. This is not easy to find

out, and LHC suggest using mileage accumulation data garnered from other data sources, such as customer survey's.

In this article, we are concerned with the same problem, that is estimation of $f_{T,M}(t,m)$. However, we attack the problem from a different direction; we choose to first estimate $f_T(t)$, the marginal distribution of T, and then estimate $f_{M|T}(m|t)$, the conditional distribution of M given T. We then have $f_{T,M}(t,m)=f_T(t)f_{M|T}(m|t)$.

The marginal density of T is much easier to estimate than that of M, since we know when each automobile was sold, and when each warranty claim (failure) is made. Note also that $f_{M|T}(m|t)$ is generally *not* the same as $f_M(m)$ (unless M and T are independent), and is definitely not to be confused with the marginal distribution of U, the mileage accumulation rate across the total population of drivers.

2. Method

Methods for estimating $f_T(t)$ are well documented in the literature for standard cases. One of the most common methods is to construct the Kaplan-Meier (K-M) estimator $\hat{S}(t)$ say (Kaplan & Meier, 1958), of the *Survivor Function*, $S(t)=\Pr(T>t)$, and then plot the estimated *Cumulative Hazard* function given by $\hat{H}(t) = \log[-\hat{S}(t)]$; for example, see Crowder, *et al.* (1991, pp 45). Since the Kaplan-Meier estimator is non-parametric, plots of $\hat{H}(t)$ against time might suggest parametric forms for $f_T(t)$, which could then be estimated more formally, e.g. via maximum likelihood. For example, if $\hat{H}(t)$ plots as a straight line, an exponential distribution would be appropriate; if it plots as a quadratic, a Weibull distribution with shape parameter 2 is appropriate. See Nelson (1972) for a comprehensive review of estimating $f_T(t)$ from plots of $\hat{H}(t)$. If no standard distribution seems appropriate, the density function for T can be obtained from the general result

$$f_T(t) = h(t)\exp\left[-\int_0^t h(u)du\right]$$

where $h(t)$, the hazard function, is the derivative of $H(t)$. The cumulative hazard can be parameterized in a flexible way, e.g. with a polynomial, the only restriction being that it must be positive and an increasing function of t.

The main problem of applying the K-M estimator in automobile warranty data is that the mileage restriction on warranty (common in the US, not so much in Europe) is that estimating the risk sets (i.e. those cars unfailed at any particular time in service), can be troublesome (see Section 3).

Estimation of $f_{M|T}(m|t)$ will generally require regression methods of some sort to model the dependence of the M distribution on T. It seems sensible to assume that $M|T \sim D_M[\theta_t]$, where D_M denotes a general probability distribution for M, indexed by a vector of parameters θ_t, which depends on t. For example, we could take $M|T$ to be distributed as Weibull with scale parameter $\theta_t=\theta_0+\theta_1 t$, for some (strictly positive) θ_0 and θ_1, and shape parameter $b_t=b$,

independent of time. It then remains to estimate θ_0, θ_1, and b, either graphically if the data set is abundant enough, or more formally using the method of maximum likelihood. Of course an initial graphical analysis would suggest parametric forms for θ_t which could be incorporated into the likelihood.

Although T is strictly speaking a continuous variable, in practice we work with it on a discrete scale, at t=1,2,3,... months. Once $f_T(t)$ and $f_{M|T}(m|t)$ have been estimated, they can be combined in the way indicated to form the joint density function for T and M.

With emphasis on warranty on automobiles as here, it then remains to evaluate $f_{T|M}(t|m)$ for various conditions on T and M. For example, assessing the likely failure percentage (the warranty exposure) for a particular failure for t=36 months (=t_w, say) and m=36,000 miles (=m_w, say), a typical warranty period for a car in the United States.

We illustrate the idea with application to a real problem in Section 4, where the mileage limit on warranty was not a problem for estimating the risk sets prior to the construction of the K-M estimator for the survivor function. Further work is needed for cases when this mileage censoring is significant, and we hope to report on this at a later date.

3. Censoring

Typically in the automotive industry, there are two warranty thresholds based on time in service and mileage accumulation of the vehicle. This leads to two types of censoring that need to be considered when estimating failure rates from warranty data. These may be summarised as;

a. specimens that are still unfailed and still under warranty with varying times in service up to t, and
b. specimens that have a mileage that exceeds the warranty threshold at time t, some of which may have failed.

As far as censorings of type a) are concerned, note that, strictly speaking, there is some information on $f_{M|T}(m|t)$ contained in these items, since some of these items will eventually fail at t, and hence contribute to the estimation of $f_{M|T}(m|t)$. However, ignoring this information only has implications for precision and not bias, since it seems reasonable to assume that the eventual realisations of failure mileage's will be a further stochastic representation of the data already observed at these fail times. This is the standard type of censoring usually present when using the Kaplan-Meier estimator.

Censorings of type b) are a little more troublesome. Because their failure/survival history is not known (they are "out of warranty"), the simplest approach is to simply treat them as though they are censored at the time their mileage exceeds the warranty threshold. Since these times are unknown, and the specimens are essentially lost to follow-up, the number of these individuals has to be inferred for various t values, and the *risk set* (those vehicles whose fail times are known to exceed t) adjusted accordingly. To do this, we need some information on mileage accumulation rates (LHC's U variable) in the vehicle population. Following the recommendation in LHC, we looked at a data source from a customer survey on the same car model, which contained information on mileage accumulation independent of the warranty data set. We found that, for a given time in service, mileage could be well represented by a log-Normal distribution, that the average mileage for this type of car was around 950 miles per month, and that the standard deviation of the log-mileage was independent of time in service, at around 0.65; in other words

$$\log_e(U)|T=t \sim N(\mu=\log_e[770t], \sigma^2=0.65^2); \quad (1)$$

because the mean of a lognormal distribution is $\exp(\mu+\frac{1}{2}\sigma^2)=950$. This result is in good agreement with LHC's own findings.

Treating these vehicles as censorings in the way described is likely to work well if specimens which are high mileage accumulators are not more prone to fail than others, which would introduce potentially serious bias. A simple way to check this is to form a plot of failure mileage *versus* failure time, and impose a line at $m=950t$, representing the average mileage accumulation. If such a plot shows a propensity of failures above the imposed line, this would indicate that the failure set are more likely to exceed the warranty threshold than the unfailed set. One way round this would be to estimate $f_T(t)$ with only those vehicles with a lower time in service.

In other cases, censorings of type b) may not be an issue at all; for example, the warranty threshold mileage may be such that it is extremely unlikely that there will any failures outside the warranty period, at least for lower times in service. This might be the case if the analysis is being done early in a vehicles life (as is the case in our worked example later), or if there is no mileage threshold for the early part of the warranty coverage (which is the case in much of Western Europe).

Other mechanisms can cause censoring issues, such as vehicles being withdrawn from use altogether as the result of an accident - we have ignored this aspect completely.

4. Application

The example here is from a population of mid-size cars in the U.S. and concerns the warranty on an exhaust emissions component. Such components are particularly interesting from the perspective of joint modelling of T and M, not least because regulatory authorities require that $\Pr(T<t_c, M<m_c)$ for critical times and mileages t_c and m_c, must be demonstrably less than p_c, a critical probability level, otherwise the authorities can demand that the automobile manufacturer initiate recall action on the entire population of vehicles for replacement of the part by one demonstrably more "reliable". Because of this, most emissions components have a warranty with $t_c=t_w$ and $m_c=m_w$.

There are 91,062 cars in the population of interest, with 1208 warranty claims recorded within the warranty period of $t_c=t_w=96$ months and $m_c=m_w=80{,}000$ miles. The oldest vehicles in the field at the time of the analysis had just over 24 months of service. Using (1) we estimated that 899 cars had a mileage in excess of m_w when we made our analysis.

Figure 1 which shows the 1208 warranty claims in the population of 91,062 cars. Although there is some evidence that the failures are occurring on high mileage accumulators, it seems unlikely that there will be disproportionate number of failures among the 899 vehicles above the threshold of 80,000 miles, so we choose not to worry too much about censorings of type b).

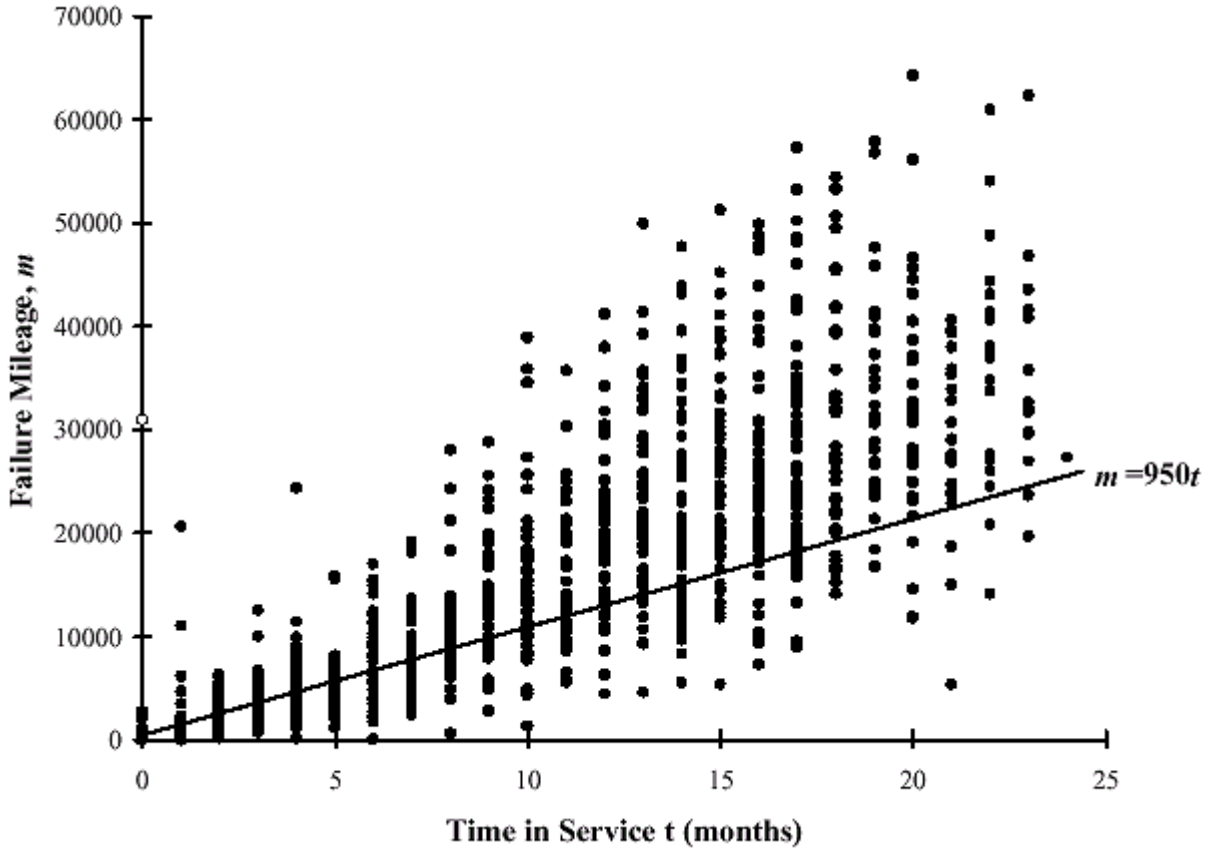

Figure 1: *Plot of failure mileages at corresponding times in service for 1208 warranty claims. The line represents the average mileage as a function of time in service (950 miles per month).*

Beginning with T, time to failure, first, the data set typically looks like that illustrated in Table 1.

Table 1: *Typical warranty data set. The n_i represent sales volumes, and the $d_{i,j}$ represent numbe failures for each months sales at increasing times in service.*

Calendar Month	Sales Volume	Time in Service (months) 1	2	3	4	5	6	7	8	9	10	11
August	n_1	$d_{1,1}$	$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{1,5}$	$d_{1,6}$	$d_{1,7}$	$d_{1,8}$	$d_{1,9}$	$d_{1,10}$	$d_{1,1}$
September	n_2	$d_{2,1}$	$d_{2,2}$	$d_{2,3}$	$d_{2,4}$	$d_{2,5}$	$d_{2,6}$	$d_{2,7}$	$d_{2,8}$	$d_{2,9}$	$d_{2,10}$	$d_{2,1}$
October	n_3	$d_{3,1}$	$d_{3,2}$	$d_{3,3}$	$d_{3,4}$	$d_{3,5}$	$d_{3,6}$	$d_{3,7}$	$d_{3,8}$	$d_{3,9}$	$d_{3,10}$	
November	n_4	$d_{4,1}$	$d_{4,2}$	$d_{4,3}$	$d_{4,4}$	$d_{4,5}$	$d_{4,6}$	$d_{4,7}$	$d_{4,8}$	$d_{4,9}$		
December	n_5	$d_{5,1}$	$d_{5,2}$	$d_{5,3}$	$d_{5,4}$	$d_{5,5}$	$d_{5,6}$	$d_{5,7}$	$d_{5,8}$			
January	n_6	$d_{6,1}$	$d_{6,2}$	$d_{6,3}$	$d_{6,4}$	$d_{6,5}$	$d_{6,6}$	$d_{6,7}$				
February	n_7	$d_{7,1}$	$d_{7,2}$	$d_{7,3}$	$d_{7,4}$	$d_{7,5}$	$d_{7,6}$					
March	n_8	$d_{8,1}$	$d_{8,2}$	$d_{8,3}$	$d_{8,4}$	$d_{8,5}$						
April	n_9	$d_{9,1}$	$d_{9,2}$	$d_{9,3}$	$d_{9,4}$							
May	n_{10}	$d_{10,1}$	$d_{10,2}$	$d_{10,3}$								
June	n_{11}	$d_{11,1}$	$d_{11,2}$									
July	n_{12}	$d_{12,1}$										

Note that as one moves down the table, less and less data is available because these vehicles have been in service for less time than those above them in the table. To obtain the Kaplan-Meier estimator of the survivor function for the entire production volume, at a particular point in calendar time, this data structure needs to be taken into account so that risk sets can be adjusted in an appropriate way through time. For example, in Table 1, at a typical time j, the risk set is $r_j = \sum_{l=1}^{12-j+1}(n_l - \Sigma_{m=0}^{j-1} d_{l,m})$, which gives the number of cars whose fail times equal or exceed j months. However, this ignores type b) censorings due to cars above the mileage threshold. A further adjustment is to use (1) and subtract from these risk sets the estimated number of vehicles which will have exceeded m_w=80,000 miles for j=1,2,3... months in turn.

Figure 2 shows the resulting Kaplan-Meier estimate of the cumulative hazard function for the data set for which we are concerned, calculated from data similar in structure to that in Table 1. As a first approximation, a straight line can be fitted to this plot, to represent an exponential distribution for $f_T(t)$. The gradient gives an estimate for the exponential parameter, $\hat{\alpha} = 0.0065 / 24 = 0.000271$.

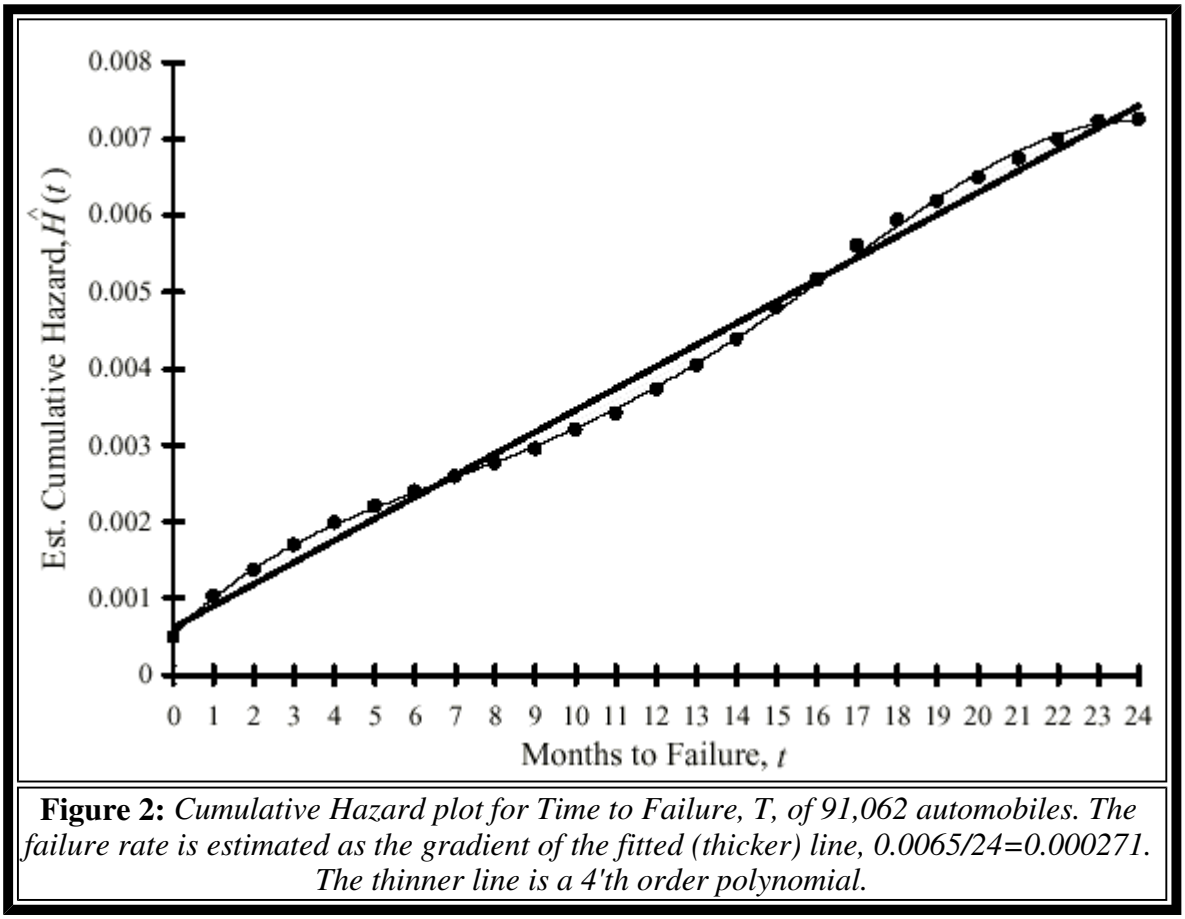

Figure 2: *Cumulative Hazard plot for Time to Failure, T, of 91,062 automobiles. The failure rate is estimated as the gradient of the fitted (thicker) line, 0.0065/24=0.000271. The thinner line is a 4'th order polynomial.*

For each of the $d_{i,j}$ failures in Table 1 recorded for the i'th production month at j months in service, the failure mileage is known from the warranty claim. We can therefore estimate $f_{M|T}(m|t)$, for t=1,2,3,... months, using these individual mileages for the $\Sigma_i d_{i,j}$ specimens in each *column* of Table 1. We assume that ignoring any failures in the set of type b) censorings will have negligible effect on these estimates, again based on the evidence in Figure 1. Any type b) censorings will have the largest effect on the higher t values.

We chose a Weibull distribution to model M, given $T=t$ (another candidate might have been the log-normal, so that M had the same distributional form as U, but the log-normal hazard function has

increasing failure rates for lower mileages, then decreasing failure rates for higher mileages, which was not considered appropriate for this problem). That is,

$$f_{M/T}(m/t) = \frac{b_t m^{b_t-1}}{\theta_t^{b_t}} \exp\left[-\left(\frac{m}{\theta_t}\right)^{b_t}\right] \qquad (2)$$

where the scale parameter, θ_t, and the shape parameter, b_t, may depend on time. For this example, there was enough data to fit separate Weibull densities for t=1,2,3,...,24 months to failure, and so we do not need to specify *a priori* parametric forms for θ_t and b_t. In cases where this is not so, regression methods using maximum likelihood to fit a proposed parametric form for θ_t and b_t will be needed (e.g. see Crowder, et al., 1991, Chapter 4).

A simple way to fit a Weibull density using graphical methods is given by Nelson (1972), which involves plotting the log of the (ranked) fail times, t_i, against the plotting position $p_i=(i-½)/n$ on a double-log scale, $\log(-\log(1-p_i))$. Of course, more formal methods such as individual maximum likelihood estimates could also be used.

Figure 3 illustrates these Weibull plots for different failure times. The gradient of each plot gives an estimate of b, and the intercept (corresponding to p=0.632) gives an estimate of θ.

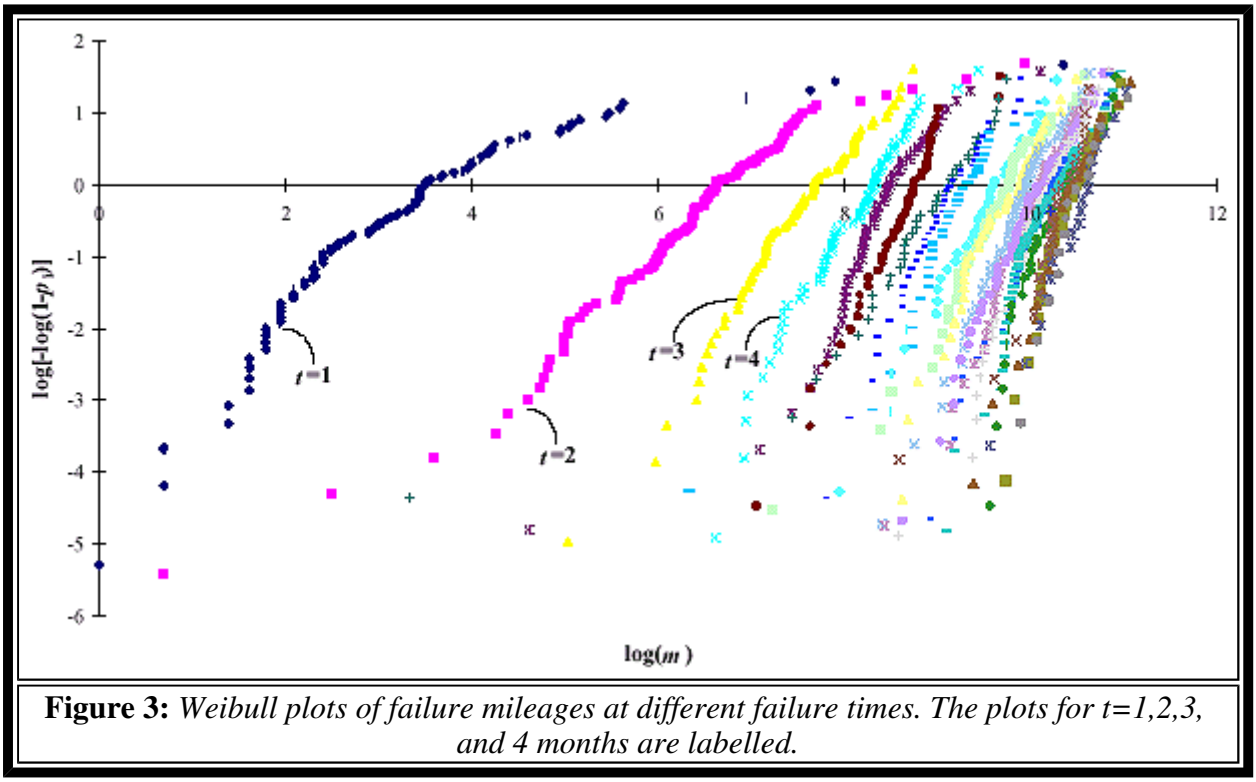

Figure 3: *Weibull plots of failure mileages at different failure times. The plots for t=1,2,3, and 4 months are labelled.*

Figures 4 and 5 show plots of θ and b, derived from the plots in Figure 3, against time. Here, we can clearly see that both of these parameters depend on time, and a linear approximation in both cases seems adequate, at least for an initial model. Appropriately, the plot for θ passes through the origin, the units of θ being miles. Also note that the shape parameter, b_t, is greater than 1 for all t, implying an increasing failure rate over mileage (conditional on T).

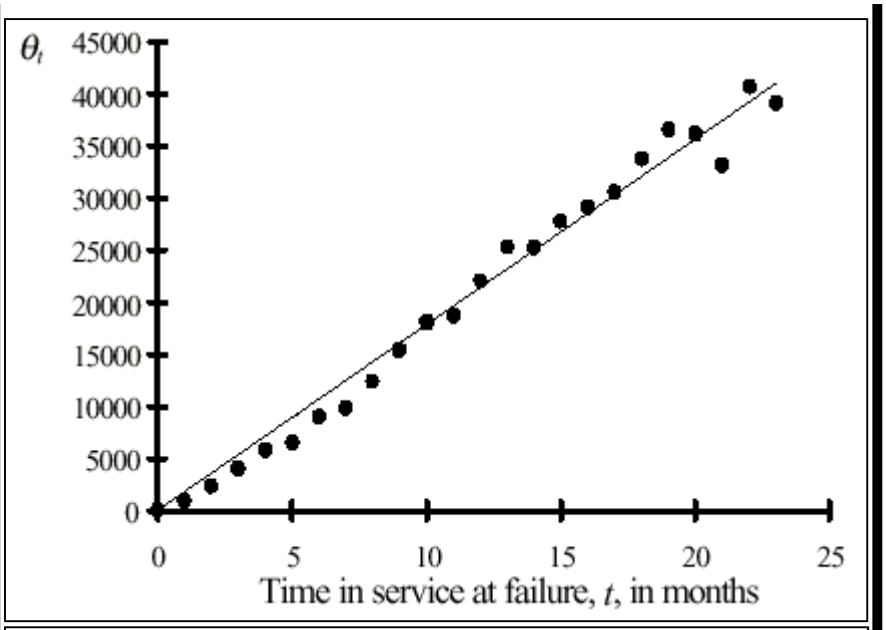

Figure 4: *Plot of fitted* θ *values versus time (in months) at failure. The line represents* $\theta_t = 1780t$.

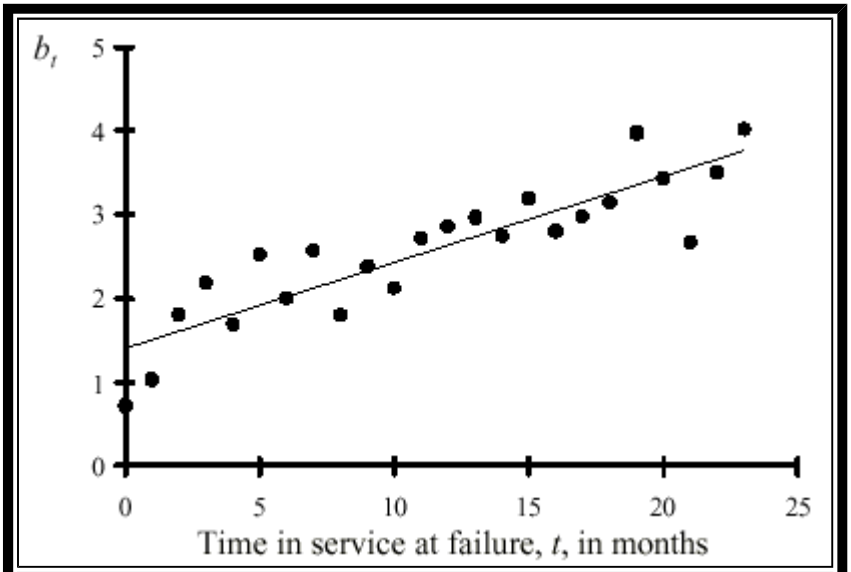

Figure 5: *Plot of fitted b values versus time (in months) at failure. The line represents* $b_t = 1.4 + 0.1t$.

The parametric forms for θ_t and b_t can be substituted into (2), to give the density function for $f_{M|T}(m|t)$. The marginal density for T is taken to be exponential from Figure 2, i.e. $f_T(t) = \alpha\exp(-\alpha t)$, with parameter estimate $\hat{\alpha} = 0.000271$; multiplying these two functions together gives our estimate for the joint density of T and M, which is illustrated in Figure 6. The resulting equation is a little intractable for analytical integration. We first tried to get an estimate of the probability p_c for critical values of t_c=96 months and m_c=80,000 months using numerical integration in the spreadsheet program used to draw Figure 6, and then verified results using specific integration software. The spreadsheet method worked very well.

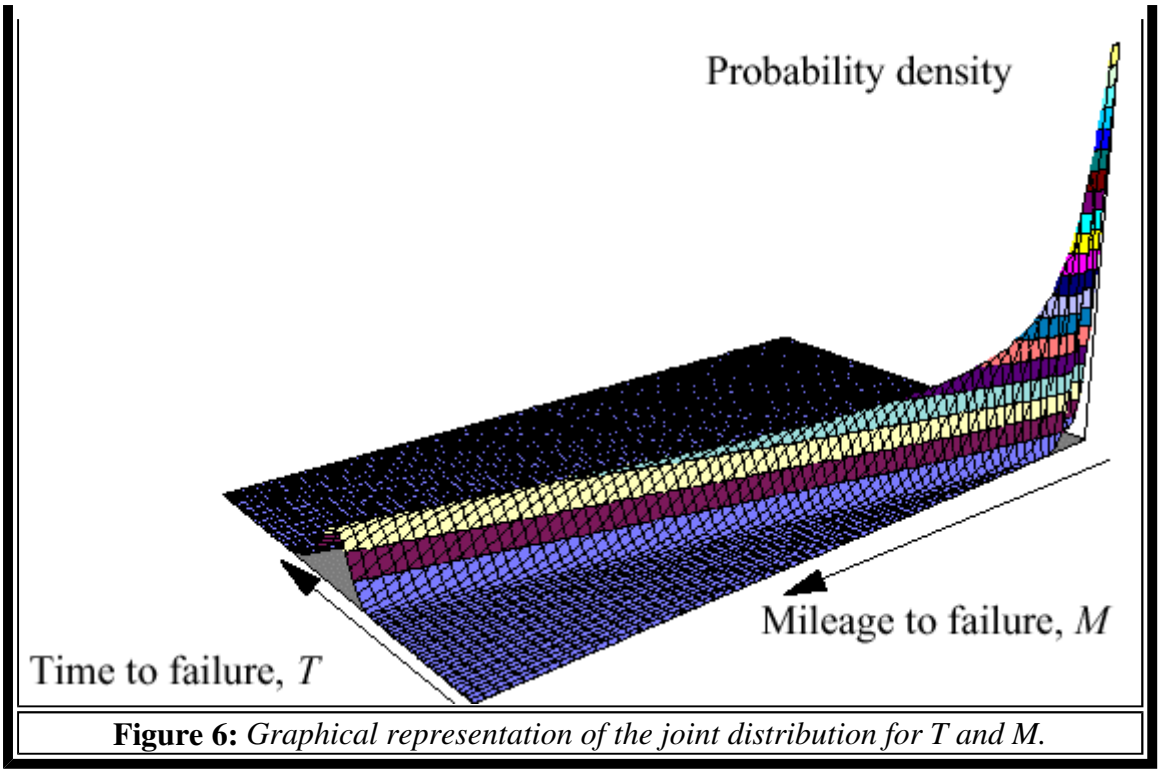

Figure 6: *Graphical representation of the joint distribution for T and M.*

Note that, if t_c and m_c are contained within the range of available data, and therefore prediction does not need extrapolation, non-parametric estimates can be used in place of parametric ones. For example, $\hat{f}_T(t)$ can be derived directly from $\hat{S}_T(t)$ the Kaplan-Meier estimator.

5. Extensions

Maximum likelihood methods can be used to obtain more formal estimates of the parameters in the distribution, once the parametric form has been suggested by the graphical analysis. For the Weibull distribution, Chapter Four of Crowder, et al. (1991) is particularly relevant for estimating the Weibull parameters when these parameters depend on time, as here. Also, it may not always be the case that data is abundant enough to empirically fit Weibull (or any other) distributions across T, and more a structured modelling approach would then be required from the outset.

We took sales dates as time zero and assumed that there was no need to model the failure rates using month of production as a covariate, since there were no manufacturing or engineering design changes during the period of this study. Kalbfleisch, Lawless, and Robinson (1991) considered the problem of lags from production to sales date in some detail.

We have not pursued extensive model checking diagnostics. Some of the methods cited in LHC can be adapted to the modelling framework outlined here. Also, extensions to include covariates in the models for $f_T(t)$ and $f_{M|T}(m|t)$ would be reasonably straightforward; for example, the failure time experience could depend on the environmental conditions under which the automobiles were being operated, and the failure mileage's might depend on in-service duty cycles. The texts by Kalbfleisch and Prentice (1980) and Lawless (1982) contain thorough treatments of including covariates in reliability models.

6. Concluding Remarks

The approach in this paper has been motivated by the need to answer (quickly) an important question posed by warranty data in the automobile industry. All the analysis contained in this paper was conducted in a spreadsheet program, hence the emphasis on empirical and graphical methods to fit the density $f_{T,M}(t,m)$; we have deliberately avoided detail on specifying and estimating models - much good discussion is contained in the LCH paper, and the references cited there, and we recommend readers study that paper in conjunction with this article.

Our main purpose in the analysis of this data set has been to simplify the approach adopted by LHC by modelling $f_T(t)$ and $f_{M|T}(m|t)$, rather than $f_M(m)$ and $f_{T|M}(t|m)$.

The analysis of automobile warranty data for reliability prediction in the field relies heavily on the *failure* date being equal to the claim date in the warranty data-base. For the failure mode on the emissions component discussed here, that was the case, since the failure necessitated an immediate visit to the dealer. For other problems, this may not be the case, and customers will often wait until the car is due for a regular maintenance check before having a problem fixed. Indeed, there is some evidence of that here - in Figure 2, the 4'th order polynomial fitted to the hazard shows turning points at around 12 & 24 months in service, around the times automobiles visit the dealerships for these routine checks.

Acknowledgments

Stephanie Sherer provided most of the data for this work and also did some initial analysis prior to the work reported on here. Ulrich Horstmann did most of the spreadsheet programming and provided valuable suggestions during the analysis. E-mail correspondence with Martin Crowder led to some improvements in this version of the paper.

References

1. M.J. Crowder, A.C. Kimber, R.L. Smith, & T.J. Sweeting. *The statistical analysis of reliability data*, Chapman & Hall, London, 1991.
2. J.D. Kalbfleisch & R.L. Prentice. *The statistical analysis of failure time data*. Wiley, New York, 1980.
3. J.D. Kalbfleish, J.F. Lawless, & J.A. Robinson. "Methods for the analysis and prediction of warranty claims". *Technometrics*, Vol 33, pp 273-286, 1991.
4. E.L. Kaplan & P. Meier. "Non-parametric estimation from incomplete observations", *Journal of the American Statistical Association*, Vol 53, pp457-481, 1958.
5. J.F. Lawless. *Statistical models and methods for lifetime data*. Wiley, New York, 1982.
6. J. Lawless, J. Hu, & J Cao. "Methods for the estimation of failure distributions and rates from automobile warranty data", *Lifetime Data Analysis*, Vol. 1, pp227-240, 1995.
7. W. Nelson. "Theory and applications of hazard plotting for censored failure data", *Technometrics*, Vol 4, pp945-966, 1972.

About the Author

Tim P. Davis
Ford Werke AG
Ford of Germany
Ford Motor Company

Last Modified *February 24, 2004*